The Clean Paper

WHAT THE PAPER SAYS · WHAT IT DOESN'T · WHY IT MATTERS

AI klinis yang tampak aman dan memperbaiki dokumentasi — tetapi tidak meningkatkan outcome pasien

Uji pragmatis dengan randomisasi klaster di 16 fasilitas layanan primer di Kenya menguji alat pendukung keputusan berbasis AI generatif yang ditambahkan ke rekam medis clinical officer. Di antara 9.691 pasien, komposit ketat kegagalan pengobatan dalam 14 hari yang dinilai ahli adalah 2,2% dengan alat versus 2,0% tanpa alat (adjusted odds ratio 0,77, 95% CI 0,55-1,08, P = 0,13) — tidak ada perbedaan signifikan. Alat tidak menunjukkan sinyal keamanan, meningkatkan dokumentasi pada semua domain yang dinilai, tidak mengubah peresepan maupun kepuasan, dan cenderung menurunkan biaya antibiotik. Ini bukan studi yang gagal; ini studi langka yang jujur — diukur pada pasien, bukan benchmark — dan menunjukkan kenyataan yang tidak glamor: alat yang tampak aman dan membantu proses belum terbukti membantu pasien dalam dua minggu, dan untuk mendeteksi manfaat seperti itu dibutuhkan uji yang jauh lebih besar.

Written by Lucio Vaglio · Cross-checked by Laura Nesso · Edited by Michele Renda · Figures by Laura Nesso and Nova Vela · AI-assisted, human-reviewed

Paper: *Generative AI-enabled clinical decision support system in primary care: a pragmatic, cluster-randomized trial*, Ambrose Agweyu, Paul Mwaniki, Vaishnavi Menon, Robert Korom, Lynda Isaaka, Conrad Wanyama, Xiaoxuan Liu & Bilal A. Mateen (and colleagues), Nature Medicine (2026) · DOI: 10.1038/s41591-026-04503-6

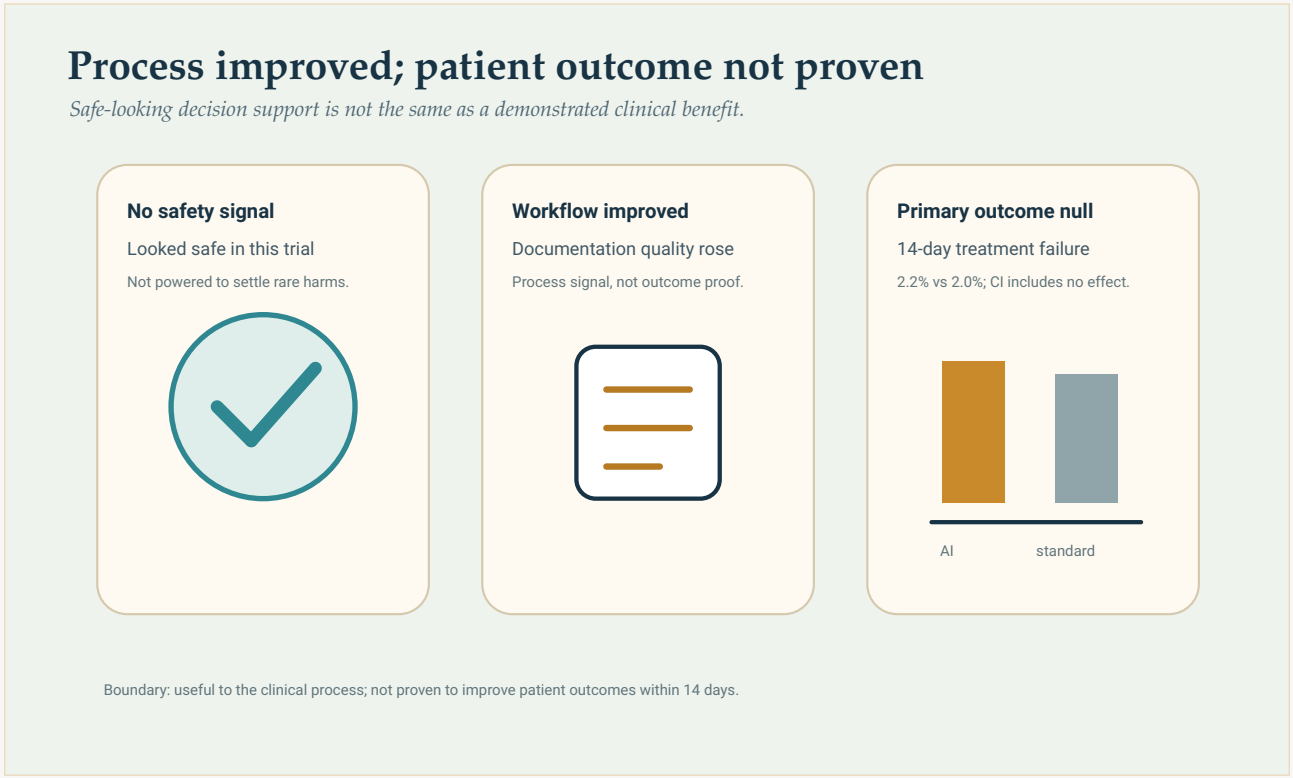
Aman dan membantu tidak sama dengan efektif

Sebagian besar judul berita tentang AI medis dibangun dari jenis studi yang keliru. Sebuah model mendapat nilai tinggi pada soal ujian, atau mengungguli dokter pada vignette kasus yang sudah dipilih rapi, lalu ceritanya seolah menulis dirinya sendiri: mesin itu siap masuk klinik.

Uji ini melakukan sesuatu yang lebih sulit dan jauh lebih jarang. Ia menempatkan alat pendukung keputusan berbasis AI generatif ke dalam layanan primer nyata, di fasilitas nyata, dengan pasien nyata, lalu menanyakan hal yang benar-benar penting: apakah pasien menjadi lebih baik?

Jawaban yang jujur adalah tidak — setidaknya tidak secara terukur dalam 14 hari. Alat tersebut tidak menunjukkan sinyal keamanan. Kualitas dokumentasi klinis membaik. Bahkan mungkin ada sedikit penurunan biaya obat. Namun kegagalan pengobatan tidak berkurang secara signifikan, dan para penulis berhati-hati menyatakan bahwa jika memang ada manfaat, kemungkinan besarnya kecil.

Itu bukan kegagalan studi. Justru itulah studi yang bekerja sebagaimana mestinya. Beginilah bukti yang bertanggung jawab tentang AI dalam kedokteran terlihat ketika diukur pada pasien, bukan pada tes pembanding.



Alat tidak menunjukkan sinyal keamanan dan meningkatkan dokumentasi klinis, tetapi luaran pasien 14 hari yang telah ditentukan sebelumnya tidak membaik secara signifikan. Membantu proses tidak sama dengan membuktikan manfaat bagi pasien.

Apa yang dilakukan para penulis

Tim menjalankan uji pragmatis dengan randomisasi klaster di **16 fasilitas layanan primer** milik jaringan kesehatan swasta Penda Health di wilayah Nairobi dan Kiambu, Kenya. Perawatan di fasilitas ini sebagian besar diberikan oleh **petugas klinis** — praktisi tingkat menengah dengan diploma tiga tahun — yang sering kali tidak memiliki akses mudah untuk berkonsultasi dengan tenaga senior.

Unit randomisasi adalah klinisi, bukan pasien. **103 petugas klinis** diacak: 52 ke kelompok intervensi dan 51 ke kelompok kontrol. Kedua kelompok menggunakan rekam medis elektronik berbasis cloud yang sama. Kelompok intervensi juga mendapat **"AI Consult"** (versi 2.0), alat pendukung keputusan yang dibangun di atas large language model **GPT-4o milik OpenAI** dan tertanam dalam rekam medis tersebut. Alat membaca informasi yang dicatat klinisi dan dapat menandai kemungkinan masalah pada diagnosis atau rencana terapi. Klinisi tetap memiliki kendali penuh: mereka dapat menerima, mengubah, atau mengabaikan sarannya.

> **Model apa yang digunakan, dan mengapa detailnya penting**
>
> Alat yang digunakan adalah **AI Consult 2.0**, yang menjalankan **GPT-4o milik OpenAI** (rilis Mei 2025), diakses melalui API komersial OpenAI dengan lisensi enterprise dan dijalankan dengan pengaturan keacakan rendah (temperature 0,1). Alat berada di dalam rekam medis elektronik khusus (EMR EasyClinic) dan diarahkan oleh system prompt yang ditulis agar selaras dengan pedoman terapi nasional Kenya; para penulis memublikasikan prompt instruksi lengkapnya. Mengapa perlu dirinci? Karena hasil ini berlaku untuk *satu sistem tertentu* — satu versi model, satu prompt, satu rekam medis, satu konfigurasi — bukan untuk "LLM dalam kedokteran" secara umum. Para penulis menyampaikan hal yang sama: mereka menyebut temuan ini sebagai *tolok ukur temporal, bukan estimasi kemampuan yang tetap*. Model yang lebih baru, prompt yang berbeda, atau klinik yang kurang terdigitalisasi semuanya dapat mengubah hasil. Mengenai independensi: OpenAI kemudian memberikan dukungan dalam bentuk barang/-jasa (kredit komputasi cloud dan panduan teknis penggunaan API), tetapi para penulis menyatakan keputusan memakai OpenAI dibuat *sebelum* tawaran itu, dan OpenAI tidak berperan dalam desain uji, pengumpulan data, analisis, maupun keputusan untuk memublikasikan hasil.

Antara 22 April dan 16 Juli 2025, **9.691 pasien** direkrut. **Luaran primer sengaja dibuat berpusat pada pasien dan ketat**: sebuah *komposit kegagalan pengobatan* dalam 14 hari setelah kunjungan yang dinilai oleh panel ahli — sekelompok klinisi, tanpa mengetahui kelompok studi, menilai apakah setiap pasien mengalami luaran buruk seperti penyakit yang tidak membaik atau justru memburuk. Uji didaftarkan sebelumnya (Pan-African Clinical Trials Registry 202502499779176).

Pilihan desain itu adalah inti studi. Mudah menunjukkan bahwa alat AI mengubah apa yang ditulis klinisi. Jauh lebih sulit — dan jauh lebih bermakna — menunjukkan bahwa alat tersebut mengubah apa yang terjadi pada pasien.

Apa yang mereka temukan

Luaran primer tidak membaik. Kegagalan pengobatan terjadi pada 102 dari 4.693 pasien (2,2%) di kelompok AI dan 94 dari 4.654 pasien (2,0%) di kelompok kontrol. Persentase mentah sedikit lebih tinggi dengan AI, tetapi setelah penyesuaian estimasi titik bergerak ke arah manfaat: **odds ratio tersesuaikan 0,77 (interval kepercayaan 95% 0,55 hingga 1,08, P = 0,13)** — tidak signifikan secara statistik. Perubahan arah antara angka mentah dan angka yang disesuaikan bukan kesalahan hitung; penyesuaian memperhitungkan perbedaan antarklaster klinisi. Bagaimanapun, interval kepercayaan dengan nyaman mencakup "tidak ada efek", sehingga manfaat tidak dapat diklaim, dan secara absolut efeknya sangat kecil.

Untuk cara sederhana membaca odds ratio, interval kepercayaan, dan nilai P secara bersamaan, lihat panduan membaca hasil klinis.

Tidak ada sinyal keamanan — dalam batas studi. Tidak ada kejadian tidak diinginkan serius yang dinilai terkait dengan alat, dan tinjauan independen tidak menemukan sinyal keamanan. Para penulis jujur mengenai batas dari kepastian tersebut: uji ini tidak memiliki daya statistik untuk mendeteksi bahaya berat yang langka, dan tidak memiliki kerangka noninferioritas atau keamanan formal yang telah ditentukan sebelumnya, sehingga tidak dapat *membuktikan* keamanan terhadap kejadian berat yang jarang.

Dokumentasi membaik. Pada 2.000 konsultasi yang ditinjau oleh ahli secara tersamar, klinisi yang menggunakan AI Consult menghasilkan dokumentasi klinis yang lebih baik pada semua domain yang dinilai — diagnosis yang dicatat, rencana terapi, dan kelengkapan secara keseluruhan.

Peresepan hampir tidak berubah. Tidak ada perbedaan signifikan dalam peresepan, termasuk penggunaan antibiotik yang benar (odds ratio tersesuaikan 0,86, 95% CI 0,48 hingga 1,55). Alat juga tidak mengubah tingkat peresepan antibiotik.

Pasien tidak merasakan perbedaan. Di antara 826 pasien yang mengisi survei kepuasan, kepuasan hampir sama di kedua kelompok, dan durasi konsultasi juga serupa.

Biaya sedikit mengarah turun. Dalam analisis yang disesuaikan, biaya terkait antibiotik lebih rendah di kelompok AI — mungkin karena pilihan obat yang lebih murah, bukan karena resep lebih sedikit — dan penghematan antibiotik per pasien tampak lebih besar daripada biaya menjalankan alat per pasien. Para penulis menandainya sebagai sinyal yang menarik, bukan kesimpulan: perhitungan total biaya kepemilikan lengkap berada di luar cakupan uji.

Ringkasan satu kalimat para penulis sendiri adalah versi paling bersih: bantuan LLM tampak aman dalam batas tersebut tetapi tidak menurunkan kegagalan pengobatan dalam 14 hari, dan jika ada manfaat, kemungkinan besarnya kecil.

Mengapa “tidak ada perbedaan signifikan” bukan berarti “tidak bekerja”

Luaran primer yang nihil mudah dibaca terlalu jauh ke salah satu arah. Ada dua hal yang mencegah cerita sederhana itu.

Pertama, **uji dirancang untuk menangkap efek yang lebih besar daripada yang ditemukan.** Outcome buruk serius dalam layanan primer jarang — sekitar 2% di sini — sehingga membedakan manfaat kecil yang nyata dari noise memerlukan jumlah peserta yang sangat besar. Perhitungan daya statistik post-hoc para penulis menunjukkan bahwa untuk mendeteksi efek sebesar yang mereka amati diperlukan **uji yang jauh lebih besar, sekitar lebih dari 100.000 pasien.** Hasil tidak signifikan dalam studi sebesar ini tidak menyingkirkan manfaat kecil yang nyata; artinya studi ini tidak mampu memastikan manfaat tersebut.

Kedua, **perbandingan antar-kelompok sebagian menjadi kabur.** Kesalahan konfigurasi sempat memberi beberapa klinisi kelompok kontrol akses ke AI Consult, dan klinisi dalam satu jaringan saling berbicara serta membawa kebiasaan melintasi batas kelompok. Kedua hal itu cenderung membuat kedua kelompok tampak lebih mirip, mendorong perbedaan nyata ke arah nol. Selain itu, jaringan tempat studi berlangsung sudah beroperasi dengan standar yang relatif tinggi, sehingga ruang untuk perbaikan oleh alat lebih kecil.

Tidak satu pun hal ini menyelamatkan judul “terobosan”. Namun artinya pembacaan yang tepat harus terkalibrasi, bukan sekadar meremehkan: pada luaran yang paling ketat dan paling relevan, alat ini tidak terbukti membantu pasien dalam dua minggu — sementara dokumentasi memang membaik secara terukur dan tidak muncul sinyal keamanan.

Apa yang *tidak* dibuktikan

- Studi ini **tidak** menunjukkan bahwa AI meningkatkan luaran pasien. Pada luaran primer 14 hari, tidak ada manfaat signifikan.
- Studi ini **tidak** menunjukkan bahwa AI tidak berguna. Estimasi titik mengarah ke manfaat, dokumentasi membaik di semua domain, dan biaya obat cenderung turun; hasil nihil masih konsisten dengan manfaat nyata yang kecil dan terlalu kecil untuk dikonfirmasi oleh uji ini.
- Studi ini **tidak** membuktikan alat aman terhadap bahaya langka. Tidak ada sinyal keamanan, tetapi uji tidak memiliki daya atau desain untuk mensertifikasi keamanan terhadap kejadian berat yang jarang.
- Studi ini **tidak** menunjukkan bahwa “AI mengalahkan dokter” atau menggantikan klinisi. Ini adalah *pendukung* keputusan; klinisi tetap memiliki otoritas penuh untuk menerima atau menolak saran.
- Hasil ini **tidak** otomatis dapat digeneralisasi. Uji berlangsung di satu jaringan swasta perkotaan di Kenya; lingkungan pedesaan, peri-urban, atau negara berpendapatan lebih tinggi dapat memberikan hasil berbeda ke arah mana pun.
- Studi ini **tidak** membuktikan penghematan biaya. Sinyal biaya menarik, tetapi bukan evaluasi

ekonomi lengkap.

Seberapa kuat buktinya?

Untuk klaim utama — tidak ada penurunan yang terbukti pada dampak buruk jangka pendek bagi pasien — buktinya kuat *dari sisi desain* dan dengan tepat rendah hati *dari sisi kesimpulan*. Uji prospektif, terdaftar sebelumnya, acak per klaster, dengan luaran komposit pada tingkat pasien yang dinilai secara tersamar, mendekati jenis bukti dunia nyata terbaik yang bisa dikumpulkan untuk alat seperti ini. Itu jauh lebih informatif daripada skor tolok ukur atau studi vignette.

Untuk temuan sekunder — dokumentasi lebih baik, peresepan tidak berubah, kepuasan serupa, biaya antibiotik lebih rendah — buktinya baik tetapi harus dibaca sebagai sekunder: sinyal pendukung, bukan judul utama, dan terkena keterbatasan kontaminasi serta satu jaringan yang sama.

Untuk keamanan, buktinya menenangkan tetapi terbatas: tidak ada sinyal yang ditemukan, namun ini bukan studi yang dirancang untuk menemukan bahaya langka.

Posisi yang paling berguna bukan “berhasil” maupun “gagal”. Yang lebih tepat: uji dunia nyata yang hati-hati menemukan bahwa alat AI ini tidak memunculkan sinyal keamanan dan memperbaiki proses perawatan, tanpa menunjukkan manfaat pada luaran pasien dalam dua minggu — dan untuk mendeteksi manfaat seperti itu dibutuhkan studi yang jauh lebih besar.

Mengapa ini penting

Perdebatan tentang AI medis sangat kekurangan bukti jenis ini. Ada ribuan makalah yang menunjukkan model memperoleh nilai tinggi pada ujian dan menyamai klinisi pada kasus yang rapi. Hanya sedikit uji besar, pragmatis, dan acak yang mengukur apakah pasien nyata benar-benar menjadi lebih baik. Ini salah satunya, dan hasilnya sampai pada kebenaran yang tidak glamor: lulus ujian tidak sama dengan membantu pasien.

Kesenjangan itu adalah inti ceritanya. Sebuah alat dapat benar-benar berguna bagi klinisi — catatan lebih jelas, biaya obat lebih rendah, sepasang “mata” tambahan — dan tetap tidak menggeser luaran pasien yang keras dalam dua minggu. Kedua hal itu dapat benar sekaligus, dan sistem kesehatan yang matang harus mampu memegang keduanya tanpa memilih fakta yang paling nyaman.

Studi ini juga menggeser beban pembuktian ke arah yang berguna. Jika perusahaan ingin mengklaim AI klinisnya meningkatkan perawatan, bukti yang relevan bukan papan peringkat. Bukti yang relevan adalah uji seperti ini, pada outcome yang penting bagi pasien — dan idealnya uji yang lebih besar, karena pelajaran jujurnya adalah manfaat kecil memerlukan studi besar agar dapat terlihat.

Singkatnya

Uji pragmatis dengan randomisasi klaster di 16 fasilitas layanan primer di Kenya menguji alat pendukung keputusan berbasis AI generatif ("AI Consult") yang ditambahkan ke rekam medis yang digunakan para petugas klinis. Di antara 9.691 pasien, komposit kegagalan pengobatan dalam 14 hari yang dinilai ahli adalah 2,2% dengan alat versus 2,0% tanpa alat (odds ratio tersesuaikan 0,77, 95% CI 0,55–1,08, P = 0,13) — tidak ada perbedaan signifikan. Alat tidak menunjukkan sinyal keamanan, meningkatkan dokumentasi klinis pada semua domain yang dinilai, tidak mengubah peresepan, tidak mengubah kepuasan pasien, dan terkait dengan biaya antibiotik yang agak lebih rendah. Para penulis menyimpulkan alat tampak aman dalam batas studi tetapi tidak mengurangi kegagalan pengobatan, dengan manfaat apa pun kemungkinan kecil; untuk mendeteksi efek sebesar yang diamati dibutuhkan uji jauh lebih besar, sekitar 100.000 pasien. Hasil ini tidak menunjukkan bahwa AI klinis meningkatkan luaran pasien, dan juga tidak menunjukkan bahwa alat tersebut tidak berguna — hanya bahwa alat yang tampak aman dan membantu proses belum terbukti membantu pasien dalam dua minggu, di satu jaringan swasta perkotaan.

Penilaian tanpa berlebihan

Apa yang ditunjukkan makalah: Dalam uji acak dunia nyata, penambahan alat pendukung keputusan berbasis LLM ke rekam medis layanan primer tidak memunculkan sinyal keamanan, meningkatkan kualitas dokumentasi klinis, tidak mengubah peresepan, dan tidak secara signifikan menurunkan komposit ketat kegagalan pengobatan pasien dalam 14 hari.

Apa yang masuk akal tetapi belum terbukti: Bahwa alat menghasilkan penurunan kecil yang nyata pada kegagalan pengobatan tetapi terlalu kecil untuk dideteksi uji ini; bahwa ia menghemat uang setelah semua biaya dihitung; bahwa dokumentasi lebih baik pada akhirnya menghasilkan perawatan yang lebih baik.

Apa yang tidak ditunjukkan: Bahwa AI klinis meningkatkan luaran pasien; bahwa alat aman terhadap kejadian langka; bahwa ia menggantikan atau mengungguli klinisi; bahwa hasil ini berlaku pada lingkungan pedesaan atau negara berpendapatan lebih tinggi; bahwa penghematan biaya sudah terbukti.

Keterbatasan utama: Studi memiliki daya untuk efek yang lebih besar daripada yang diamati (outcome langka membutuhkan sampel sangat besar); kesalahan konfigurasi mencemari kelompok kontrol dan kebiasaan dalam jaringan bersama mengaburkan perbandingan, keduanya bias ke arah tidak ada perbedaan; satu jaringan swasta perkotaan dengan standar yang sudah tinggi; tidak ada kerangka noninferioritas atau keamanan yang telah ditentukan sebelumnya; horizon hanya 14 hari.

Seberapa besar keyakinan yang sebaiknya dimiliki pembaca umum? Tinggi bahwa alat tidak menunjukkan sinyal keamanan dalam uji ini dan meningkatkan dokumentasi. Tinggi bahwa alat tidak terbukti meningkatkan luaran pasien 14 hari di sini. Rendah untuk klaim bahwa alat "berhasil" atau "gagal" sebagai intervensi yang memberi manfaat pada pasien — pertanyaan itu benar-benar belum selesai dan membutuhkan studi jauh lebih besar. Sikap yang tepat: hasil dunia nyata yang hati-hati

tentang alat yang tidak memunculkan sinyal keamanan dan membantu proses perawatan, bukan vonis bahwa AI mengubah — atau merusak — layanan primer.

Sumber

Agweyu et al., Nature Medicine (2026)
https://doi.org/10.1038/s41591-026-04503-6

Pan-African Clinical Trials Registry, registration 202502499779176
https://pactr.samrc.ac.za/

EurekAlert / PATH press release on the trial (2026)
https://www.eurekalert.org/news-releases/1133583